Assignment-based Subjective Answers

1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

Answer :- We have Categorical variables like – Working Day, Months, Season, Weekdays, Weathersit, Year, Holiday, dteday.

For Our Model –

- WorkingDay has a positive slope. Increase in 1 unit of WorkingDay increases the cnt value by 0.0573
- Month – February has a Negative slope. Increase in 1 unit of February decreases the cnt value by 0.0673

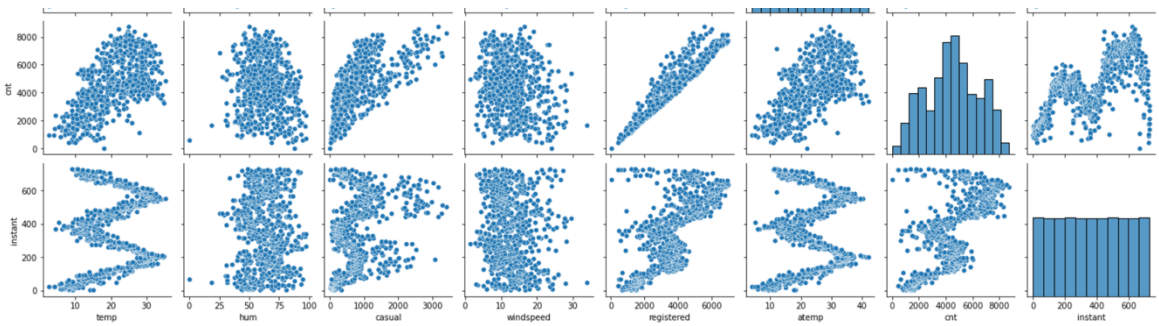
We have many categorical variables that decided the value of the dependent variable. Some of them have positive effect, Some of them have Negative Effect.

2. Why is it important to use **drop_first=True** during dummy variable creation?

Answer :- If you don't set drop_first=True, it will create n columns , thus the variables will be highly correlated, that means the first variable can be predicted from the other variables with high precision. This can lead to problems with multicollinearity in some statistical models, such as linear regression, where multicollinearity can lead to unstable and unreliable parameter estimates.

By setting drop_first=True, you drop the first column/variable and only n-1 columns are created which means it solves the problem of multicollinearity in this case.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?



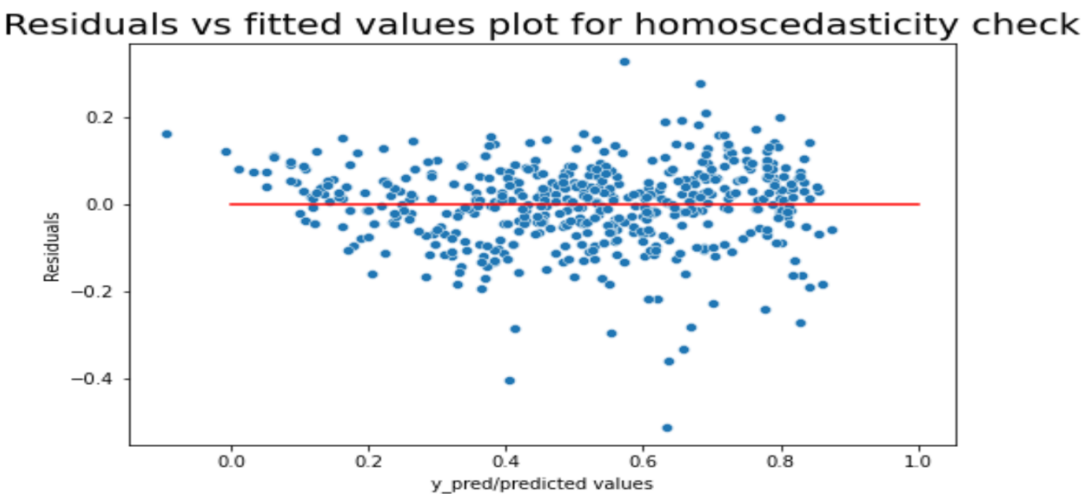
Answer :- With the Target Variable ('cnt'), the highest correlation is found between "casual", "registered" variables because the "cnt" variable is the addition of both "casual" and "registered" variables.

Apart from "casual", "registered" variables, "temp", "atemp" also have a decent positive correlation with "cnt" variable.

4. How did you validate the assumptions of Linear Regression after building the model on the training set?

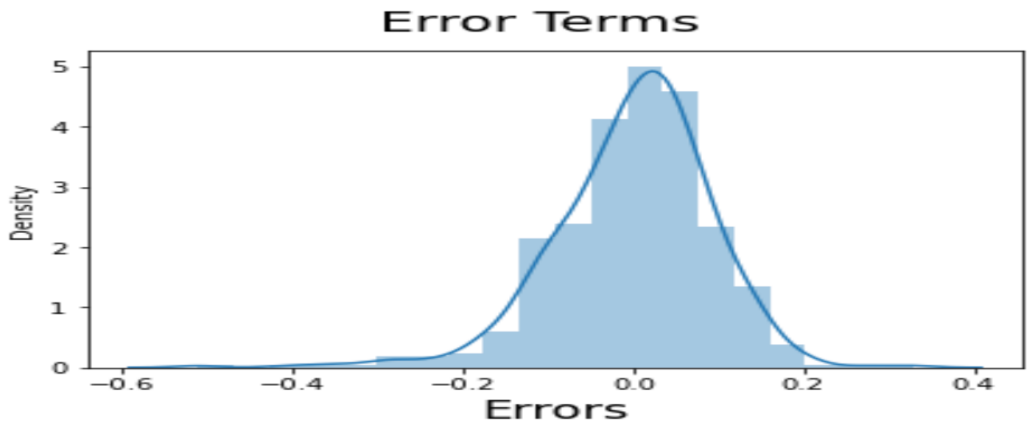
Answer :- We can use

- MultiCollinearity
    - We can use VIF values to check this
- Homoscedasticity



    - We can use Scatterplot of residuals. Based on spread we can visualize the Homoscedasticity.

- Normality of residuals



    - We can use a histogram and check if it rough bell shaped

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

Answer :- Based on My model the Top 3 Features are :-

- Due (Derieved Variable from atemp, hum)
- Year
- Weekday – Saturday

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Answer :-

Linear regression is a supervised learning algorithm used for predicting a continuous outcome variable (y) based on one or more predictor variables (x). The goal of linear regression is to find the best-fitting straight line through the data points.

The basic linear regression model is represented by the following equation:

$$y = \beta0 + \beta1x1 + \beta2x2 + ... + \beta n*xn + c$$

where y is the outcome variable, x1, x2, ..., xn are the predictor variables, β0, β1, β2, ..., βn are the coefficients of the predictors and c is the error term. The coefficients represent the change in the outcome variable for a one-unit change in the predictor variable, while holding all other predictor variables constant.

Here are the main steps of the linear regression algorithm:

- **Data preparation**: This involves cleaning and preparing the data by handling missing values, outliers and transforming variables as necessary.
- **Model Building**: This involves creating a linear regression model using the prepared data. This typically involves the following steps:

  a. **Selection of Predictors**: This involves selecting which predictor variables to include in the model. This can be done through techniques such as forward selection, backward elimination, stepwise regression, and others.

  b. **Estimation of coefficients**: Once the predictors are selected, the algorithm estimates the coefficients of the predictors in the model. This can be done through different techniques, such as ordinary least squares (OLS), gradient descent, and others. OLS is the most common method used in practice and it is done by finding the values of the coefficients that minimize the sum of the squared differences between the predicted and actual values of the outcome variable.

- **Model Evaluation**: After the coefficients are estimated, the linear regression model is evaluated to assess its performance. This typically involves assessing the model's

goodness of fit and checking the assumptions of linear regression. The most commonly used statistical measure for goodness of fit is the R-squared value.

- **Prediction**: Once the model is built and evaluated, it can be used to make predictions on new data. This is done by plugging in the values of the predictor variables into the equation of the model and solving for the outcome variable.

It's important to note that Linear regression is sensitive to the presence of outliers and multicollinearity between predictor variables. This can affect the coefficient estimates and predictions of the model. Therefore, it is important to identify and treat them before building the model.

2. Explain the Anscombe's quartet in detail ?

Answer :- Anscombe's quartet is a set of four datasets that were created by Francis Anscombe in 1973 to demonstrate the importance of visualizing data before analyzing it. The four datasets have almost identical statistical properties, but they have very different visual properties when plotted.

Each dataset consists of eleven (x, y) pairs:

- The first dataset has a clear linear relationship between x and y, with a correlation of 0.816.
- The second dataset has a non-linear relationship between x and y, but the correlation between x and y is still 0.816.
- The third dataset has an almost perfect linear relationship between x and y, but there is an outlier that greatly influences the correlation, which is 0.816.
- The fourth dataset has a linear relationship between x and y, but it has a lower correlation of 0.816 and a much higher variance in the y-values for a given x-value.

The point that Anscombe's quartet illustrate is that a single summary statistic such as the mean or correlation coefficient can give a misleading impression of the relationship between two variables, and it's crucial to visualize the data before making any assumptions or decisions. It's also a reminder that it's critical to look beyond simple summary statistics and dig deeper into the data to fully understand it.

3. What is Pearson's R?

Answer :- Pearson's R, also known as Pearson's correlation coefficient, is a measure of the linear association between two continuous variables. It's a value between -1 and 1 that indicates the strength and direction of the correlation. A value of -1 indicates a perfect negative correlation, a value of 1 indicates a perfect positive correlation, and a value of 0 indicates no correlation.

Pearson's R is calculated using the following formula:

R = (n∑xy - (∑x)(∑y)) / √((n∑x^2 - (∑x)^2)(n∑y^2 - (∑y)^2))

Where:

- n is the sample size
- ∑x and ∑y are the sums of the x and y values, respectively
- ∑xy is the sum of the product of x and y
- ∑x^2 and ∑y^2 are the sums of the squares of the x and y values, respectively.

It is important to notice that Pearson's R is only appropriate for measuring the linear relationship between two variables, it doesn't work for non-linear relationships.

Additionally, Pearson's R is not robust to outliers and is sensitive to the presence of nonnormality.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling?

Answer :-

Scaling is a pre-processing technique used to adjust the values of a variable so that they are within a specific range. The main purpose of scaling is to ensure that all the variables in a dataset are on the same scale, which is necessary for many machine learning algorithms. This is because some algorithms, such as distance-based algorithms, are sensitive to the scale of the variables, and using variables with different scales can skew the results.

There are several scaling techniques that can be used, the most common are:

- **Min-Max Scaling**: Also known as normalization, it scales the values of a variable between 0 and 1. This is done by subtracting the minimum value of the variable from each value and then dividing by the range (maximum value minus minimum value) of the variable.
- **Standardization**: Standardization, also known as z-score normalization, scales the values of a variable to have a mean of zero and a standard deviation of one. This is done by subtracting the mean of the variable from each value and then dividing by the standard deviation of the variable.

It is common practice to normalize continuous variables to scale them between 0 and 1 while standardizing variables by converting them into z-scores. But it's important to note that it's important to use the appropriate scaling technique depending on the dataset and the machine learning algorithm you're using. It is also important to note that, normalizing and standardizing are not always required, and it depends on the specific use case and the algorithm you're using.

In summary, Scaling is an important pre-processing step in machine learning because it can help to improve the performance of many algorithms by putting the variables on the same scale. Normalization scales a variable between 0 and 1, whereas standardization scales a variable to have a mean of zero and a standard deviation of one.

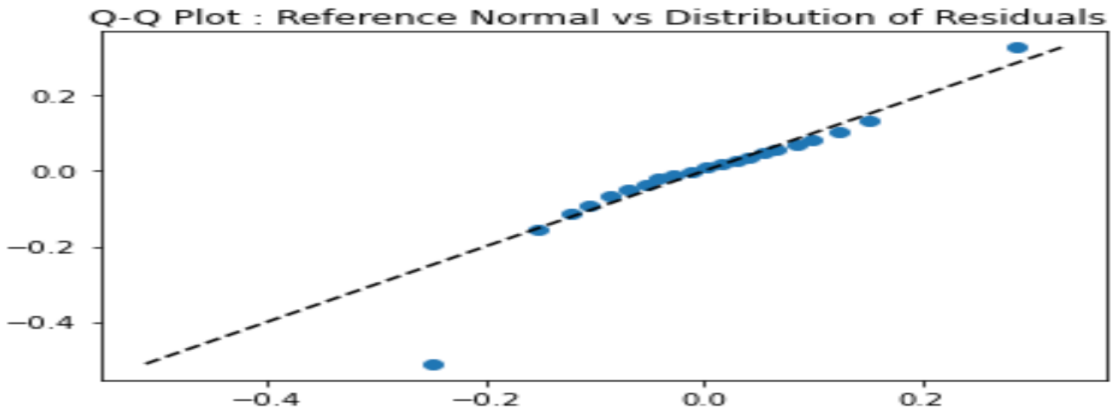
5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?

Answer :- variables in a multiple regression model. It calculates how much the variance of an estimated regression coefficient is increased due to multicollinearity. A VIF of 1 means there is no multicollinearity, whereas a VIF greater than 1 indicates that there is multicollinearity. The higher the VIF value, the stronger the multicollinearity.

A VIF value of infinite (or undefined) can occur when two or more predictor variables are perfectly correlated, also called perfect multicollinearity. This perfect multicollinearity leads to a singularity of the X'X matrix of the linear regression, which is the matrix that contains the inner product of the design matrix X. This singularity results in a division by zero or infinite values when calculating the VIF.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression ?

Answer:-



A Q-Q plot, also known as a quantile-quantile plot, is a graphical method for assessing whether a set of data comes from a specified distribution. The plot compares the quantiles of the data with the quantiles of a theoretical distribution, such as a normal distribution.

In the context of linear regression, a Q-Q plot is used to check the normality assumption of the errors (residuals) of the model. One of the assumptions of linear regression is that the errors are normally distributed. If this assumption is met, the residuals will be approximately normally distributed with a mean of zero and a constant variance.

A Q-Q plot can be used to check this assumption by plotting the residuals against a normal distribution. If the residuals follow a normal distribution, the points on the plot will fall approximately along a straight line. If the residuals are not normally distributed, the points on the plot will deviate significantly from a straight line.

A Q-Q plot is a quick and easy way to visually assess whether the normality assumption is met, and helps to detect any violations of this assumption. If the assumption is violated, then it may indicate that the model is not appropriate for the data and other methods should be considered. Alternatively, transforming the data or using a different model with different assumptions could also be considered

It's worth noting that, Q-Q plot is not the only way to check the normality of residuals, histogram, and Normality tests like Anderson-Darling, Shapiro-Wilk are also used to confirm the normality of residuals.